DOCUMENT RESUME

ED 295 982 TM 011 807

AUTHOR Kirisci, Levent; Hsu, Tse-Chi
TITLE A Predictive Analysis Approach to Adaptive Testing.
PUB DATE Apr 88
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Bayesian Statistics; Comparative Analysis; Item Analysis; Maximum Likelihood Statistics; Predictive Measurement; *Predictive Validity; Statistical Analysis; *Test Items
IDENTIFIERS *Predictive Analysis Approach

ABSTRACT

The predictive analysis approach to adaptive testing originated in the idea of statistical predictive analysis suggested by J. Aitchison and I.R. Dunsmore (1975). The adaptive testing model proposed is based on parameter-free predictive distribution. Aitchison and Dunsmore define statistical prediction analysis as the use of data obtained from an informative experiment in the past to make some reasonable statement about the outcome of the future experiment. Use of the approach's predictive density function and item selection procedure and terminating criteria is discussed. A small-scale exploration study compared the approach with A. Wald's sequential probability ratio test (1947), M. F. Lord's flexi-level test (1971), R. J. Owen's Bayesian strategy (1975), and F. C. Samejima's maximum likelihood strategy (1977). The various approaches could not be placed on equal base in terms of data used. Results indicate that: (1) final predictive probabilities were significantly correlated with total scores; (2) the predicted adaptive testing performed better than sequential probability testing and almost as well as the Bayesian strategy in the area of mastery classification; (3) the benefit of adaptive testing could not be demonstrated in the area of the number of test items required; and (4) there was no effect on the number of misclassifications of students when different priors were used in predictive testing. Three tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295982.

A Predictive Analysis Approach To Adaptive Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LEVENT KIRISCI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Levent Kirisci

and

Tse-Chi Hsu

University of Pittsburgh

TM011807

Paper presented at the annual meeting of AERA. New Orleans, LA., April, 1988.

BEST COPY AVAILABLE

INTRODUCTION

With the recent developments in item response theory and computer technology, the conditions required for the implementation of tailored testing into practice seem mature. Yet the implementations of tailored testing are confined to the situations where only large sample sizes can be obtained to calibrate items because item response theory requires large sample sizes for parameter estimations.

The purpose of the study was to develop an adaptive testing model based on parameter-free predictive distribution. In addition to the derivation of a predictive distribution, item selection strategies and terminating criteria were obtained. The feasibility of the model was also investigated by comparing its performance with the performance of Lord's flexilevel test (Lord, 1971), Wald's sequential probability ratio test (Sprt, Wald, 1947), Owen's Bayesian item selection strategy (Owen, 1975), and maximum likelihood item selection strategy (Samejima, 1977). The performance in adaptive testing was simulated using actual data obtained from a paper and pencil test.

THE MODEL

The predictive analysis approach to adaptive testing is originated from the idea of statistical predictive analysis suggested by Aitchison and Dunsmore (1975). The statistical predictive analysis is composed of two parts: Informative experiment E and future experiment F. An informative experiment E is an experiment which is performed in the past and its typical outcome is denoted by x. In the same manner, a future experiment F is an experiment which is carried out in the future and its typical outcome its denoted by y.

Aitchison and Dunsmore define statistical prediction analysis as the use of data, which are obtained from an informative experiment E in the past, to make some reasonable statement about the outcome of the future experiment F. This analysis contains two assumptions:

a) The probability distributions which describes informative experiment E and the future experiment F have the same unknown parameter space (trait).

b) For a given trait, the experiments E and F are independent

Predictive Density Function

Let ϕ be the unknown trait to be measured and $f(\phi)$ be the prior density of ϕ. Let $f(.|\phi)$ be the probability density function of x, which

is the typical outcome of the informative experiment E. Hence, the posterior density is proportional to $f(\phi|x) \propto f(x|\phi)f(\phi)$.

The predictive density function of y is obtained from the posterior distribution. Let $f(.|\phi)$ be the density function of future outcome y. By taking into consideration of Bayesian approach to predictive problems, the distribution of future outcome y given informative experiment x is defined by

$$f(y|x)=\int f(y|\phi)f(\phi|x)d\phi$$

which is called the predictive density function.

In the above derivation, it is assumed that x and y are continuous variables. The same derivation technique is applicable for discrete variables. As seen, the predictive density function does not involve parameter ϕ (trait). Yet it is possible to make inference about the magnitude of the future observations for the same trait.

Item Selection Procedure

In order to find the most appropriate item to administer to an examinee, two predictive probability functions have to be specified. One with a prior belief describes the examinee's ability level and the other with a prior belief represents the difficulty level of the item.

If a beta distribution is used to represent the prior belief of an examinee's ability level and a binomial distribution is specified for an informative function, the following predictive distribution (beta binomial) for an ability level is obtained by appliying the steps which are described in the previous section

$$f_a(y|x)=\binom{N}{y}\frac{\Gamma(\alpha+\beta)\,\Gamma(y+\alpha)\,\Gamma(N+\beta-y)}{\Gamma(\alpha)\,\Gamma(\beta)\,\Gamma(N+\alpha+\beta)}, \quad y=0, 1, \ldots, N \qquad [1]$$

where $\alpha = x+g$, $\beta=n+h-x$. The parameters of prior distribution are $g>0$ and $h>0$, where g and h are called a location parameter and a scale parameter, respectively. The sample size of an informative experiment for the number of items already administered is n and for the number of items to be administered in the future is N. The numbers of correct answers in the past experiment and in the future experiment are represented by x and y respectively.

Another predictive distribution based on the prior belief of item difficulty b can also be represented by the formula [1] except the values for the location and scale parameters of the prior distribution will be different. This predictive distribution is designated as $f_b(y|x)$. To obtain the probability of answering next

item correct, given item difficulty b and x number of items correct in the past, the proportionality of $f(y=1|b,x)$ to $f(b|x,y=1)f_B(y=1|x)$ is used, where $f(b|x,y=1)$ is the posterior probability of item difficulty given past and future information of an examinee. To find the most appropriate item to administer to an examinee the following criterion is considered

$$\min_b |f_a(y=1|x)-(1-f(y=1|b,x))|$$

The above criterion is constructed by considering the almost perfect correlation between item difficulty b and $f(y=1|b,x)$, and also the negative correlation between $f_a(y=1|x)$ and $f(y=1|b,x)$. According to the above criteria, the most appropriate item to be administered is the one whose item difficulty matches to his/her ability level.

Teminating Criteria

After a minimum number of items is administered, a decision has to be made by choosing one of the actions

a_1 = examinee is a non-master,

a_2 = no decision can be made, continue testing,

a_3 = examinee is a master.

To decide when to terminate the testing, a predictive distribution, which is based on prior belief of an examinee's ability level and a utility function are employed. The proportions of the non-mastery (R_1), undecided (R_2), and mastery (R_3) regions are designated in advance such as $\delta_1,\delta_2,\delta_3$, respectively. To make a decision, the likely number of correct answers in future which may lie in the regions are determined as follows: $k_1=\delta_1 \times N$, $k_2=\delta_2 \times N$, $k_3=\delta_3 \times N$, where N is the number of items to be administered in future and k_1, k_2, and k_3 are assumed to be the closest integer values. If the model predicts an examinee can answer only 0, 1, or upto k_1-1 items correct out of the remaining N items, the examinee is placed in the non-mastery region (R_1). If the model predicts an examinee can answer k_1 through N-k_3 items correct, the examinee is placed in the undecided region (R_2). If more than N-k_3 items correct is predicted, the examinee is placed in the mastery region (R_3). Thus, the sum of the proportions of the regions equals to 1, $\delta_1+\delta_2+\delta_3=1$, and $k_1+k_2+k_3=N$. To figure out what are the proper values for $\delta_1,\delta_2,\delta_3$, one may consider the proportions of answering the remaining N items correct. If $\delta_1,\delta_2,\delta_3$ are 0.3, 0.4, 0.3, respectively, it implies that if the examinee can answer only less than 0.3 of the

remaining items correct, he/she will be placed in the non-mastery region (R_1). If the examinee can answer 0.3 or more but less than or equal to 0.7, which is the sum of 1 and 2, he/she should be placed in the undecided region (R_2).

Let us define utility function for action a_i:

$$u(a_i, n) = \begin{cases} 1 & \text{if } y \in R_i \\ 0 & \text{otherwise} \end{cases} \qquad i=1, 2, 3.$$

Then, the terminating criterion is defined as the choice of the action a_i which gives maximum utility, $\max(\sum_{y \in R_i} u(a_i, n)f(y|x), i=1, 2, 3)$ (Aitchison and Dunsmore, 1975, Ch.8). After simplifying the above criteria, the following is obtained

$$\max(p_1=\sum_{y \in R_1} f(y|x),\ p_2=\sum_{y \in R_2} f(y|x),\ p_3=\sum_{y \in R_3} f(y|x)) \text{ or } p_3=1-p_1-p_2).$$

Then, the decision is simply the choice of the maximum probability p_i which is calculated over the region R_i.

To make a decision for any examinee that is still in undecided region (R_2) after reaching the maximum number of items to be administered three approaches were employed in this study. These three approaches are listed in order of preference, (a) comparisons of final p_1 and p_3 values, (b) comparisons of his/her final predictive probability with others having the similar predictive probabilities, and (c) comparisons of his/her number of correct answers in percentage with those of other examinees.

COMPARISONS WITH OTHER STRATEGIES

To investigate the feasibility of the predictive adaptive testing strategy, a small scale exploration study was made to compare the performance of the strategy with the performance of Lord's flexilevel test (Lord, 1971), Wald's sequential probability ratio test (Wald, 1947), Owen's Bayesian strategy (Owen, 1975), and maximum likelihood strategy (Samejima, 1977). Since the strategies could not be placed on equal base in terms of data employed, this was only a gross comparison to assess whether the predictive testing strategy is worthy of further investigations.

The comparison was made in terms of the answers to the following questions: (a) What is the relationship between total test scores and the predicted probability, estimated ability, or proportion

of correct obtained from the adaptive tests? (b) What are the proportion of misclassification into mastery or non-mastery by the adaptive tests in comparing with an arbitrary cut-off score of the total test? (c) What are the minimum number of items required for the adaptive testing decisions? (d) For strategies involving prior, what are the effect of different prior on the predictive or estimation of ability?

Table 1: Specification of strategies used in the comparison

	Predictive	Flexilevel	Wald	Bayesian	Maximum
Minimum no. of items	7	7	7	7	7
Maximum no. of items	23	23	23	23	23
Difficulty	Traditional from sample	Traditional from sample	NA	LOGIST	LOGISTS
Disrimina.	NA	NA	NA	LOGIST	LOGISTS
Guessing	NA	NA	NA	LOGIST	LOGISTS
	Beta distr.			Normal distr.	
Prior High	g=2, h=1	NA	NA	M=.5, S.D=1	NA
Prior Mid.	g=2, h=2	NA	NA	M=0, S.D=1	NA
Prior Low	g=1, h=2	NA	NA	M=-.5 S.D=1	NA
Mastery regions				Ability est	Ability est.
Mastery	δ_3=0.3	NA	0.8, 0.7	0.65 or higher	0.65 or higher
Undecided	δ_2=0.4	NA			
Non-mast.	δ_1=0.3	NA	0.5, 0.3	Below 0.65	Below 0.65
α	NA	NA	0.05	NA	NA
ß	NA	NA	0.05	NA	NA
Termination Criteria	Max(p_1,p_3)	NA		Error var 0.08	Test info 12

The data for this comparison were obtained from Form A of college math placement test. This 45-item test was administered 800 students registered for math courses. It was a part of field testing of math placement test items for developing an computerized adaptive placement test based on item response theory. Thus, the estimates of parameters, difficulty, discriminating, and guessing, are available for these 45 items (Hsu & Shermis, 1987).

In order to compare the performance of five strategies, the adaptive portion of the study was simulated. In other words, the items were administered one at a time. But the response for each item is based on the examinee's response on the answer sheet. Response data from 50 subjets were randomly selected for this comparison.

Specifications for each strategy used in the comparisons are summarized in Table 1. Several notations are in order. The maximum number of items administered was set at 23 because for a 45-item test, 23 items were required by the flexilevel strategy. Three different priors were used for the predictive strategy and Bayesian strategy. Although they are based on different distributions, they are approximately equivalent. Two set of mastery and non-mastery criteria (0.8, 0.5 and 0.7, 0.3) were employed for Wald's Sprt. They are identified as $Sprt_1$ and $Sprt_2$, respectively. Since these sets of criteria cannot be related to the mastery regions specified for the predictive strategy, the comparisons between these strategies should be interpreted with caution.

Simulation results for all strategies were compared with the results of the complete test in Table 2. Students whose complete test scores in percentage were 65 or more were assigned to the mastery group. Eleven students were classified into the mastery group and 39 were classified into the non-mastery group. By assuming that all 50 students were in the low ability level, for the predictive strategy, on the average 20 items were administered. As a result 9 out 11 students remained the mastery group. For medium ability assumption, there were 7 master students. The average number of items administered were 21. If all the students were assumed to have high ability level, 13 of them were assigned into the mastery group. However, only 10 of the 13 students were correctly classified. On the average 20 items were administered under the assumption of high ability level. The number of items used in testing all three different priors varied between 23 and 7.

In self-scoring flexilevel test, students were assigned into mastery categories based on the percent of correct answering the 23

items administered. Any student whose score in percentage was above 65 was assigned into the mastery category.

Wald's sequential ratio test were used twice with different mastery and non-mastery proportions. Twelve students were assigned into the mastery category with criteria of 0.80 and 0.50 ($Sprt_1$), of which, 8 were correctly classified. Among 25 students in the category of mastery when using 0.70 and 0.30 ($Sprt_2$), only 10 were correctly classified. The average number of items administered was 12 for $Sprt_1$ and 10 for $Sprt_2$. The number of items used for these two tests ranged between 21 and 3.

Table 2: Comparisons of average number of items administered and classification of students according to the total test scores

Adaptive Testing Strategies	Average Number (and st. dev.) of the Items Administered		No. of Students Correctly Classified as Master, Non-Master	No. of Students Misclassified as Master, Non-master (phi-coeff.)
Complete test	45	(0.00)	11, 39	0, 0 (1.00)
Pred.(Low)	19.86	(5.35)	11, 39	2, 2 (0.77)
Pred.(Med)	21.40	(4.84)	7, 43	0, 4 (0.76)
Pred.(High)	20.36	(5.17)	13, 37	3, 1 (0.79)
Flex.	23.00	(0.00)	8, 42	0, 3 (0.82)
$Sprt_1$	11.70	(6.86)	12, 38	4, 3 (0.61)
$Sprt_2$	9.86	(5.79)	25, 25	15, 1 (0.43)
Max.	22.78	(0.93)	13, 37	2, 0 (0.89)
Bayes(Low)	22.80	(0.76)	6, 44	1, 6 (0.55)
Bayes(Med)	22.28	(1.65)	8, 42	0, 3 (0.82)
Bayes(High)	20.86	(2.72)	12, 38	3, 2 (0.72)

Maximum likelihood decision strategy had the smallest total number of misclassified students into mastery category and non-mastery category. This strategy assigned 13 students into mastery category. Two of the 13 students were misclassified. Bayesian decision strategy assuming low prior ability assigned only 6 students into mastery category and only one incorrectly classified. Number of students assigned into mastery category increased when medium or

high prior ability belief were assumed. For Bayesian decision startegy with medium prior ability belief, the number of mastery students were 8 and they were all correctly classified. For high ability level assumption the number of students in mastery category was 12. Three of the 12 students were misclassified into the mastery category. The average number of items used by maximum and Bayesian strategies were 23 and 22, respectively.

Table 3 presents the correlations between total test scores (total) and the number of correct scores in percentage obtained from flexilevel test, $Sprt_1$ and $Sprt_2$. For maximum likelihood and Bayesian decision starategies correlations were computed between total score and the obtained estimated ability. For the predictive test, correlations were computed between total test scores and final predictive probabilities.

Table 3: Correlations between total test scores, the final estimate of ability scores, predictive probabilities or percentage correct scores

	Total	$Pred_L$	$Pred_M$	$Pred_H$	Flex	$Sprt_1$	$Sprt_2$	Max	Bay_L	Bay_M	Bay_H
Total	1.00	0.82	0.87	0.77	0.89	0.63	0.71	0.76	0.87	0.88	0.86
$Pred_L$		1.00	0.89	0.81	0.91	0.47	0.65	0.64	0.63	0.67	0.70
$Pred_M$			1.00	0.84	0.91	0.51	0.62	0.66	0.69	0.70	0.74
$Pred_H$				1.00	0.84	0.52	0.68	0.59	0.62	0.64	0.66
Flex					1.00	0.54	0.62	0.66	0.72	0.77	0.76
$Sprt_1$						1.00	0.89	0.41	0.51	0.54	0.49
$Sprt_2$							1.00	0.47	0.55	0.57	0.53
Max								1.00	0.75	0.78	0.76
$Bayes_L$									1.00	0.96	0.85
$Bayes_M$										1.00	0.89
$Bayes_H$											1.00

The correlations between total test scores and pred(low), pred(med), pred(high) are highly comparable with those of maximum likelihood and Bayesian strategies. The correlation coefficients between predictive tests and complete test scores are

higher than the correlations between the complete test scores and both Wald's sequential tests, and the correlation between the complete test scores and the maximum likelihood strategy. It should be noted that the average total test score of 50 students in percentage was 47 and the median score was 42. Therefore, low ability or medium ability
assumptions were more appropriate than high ability assumption. This is probably the reason why the correlations between the complete test and predictive test assuming high ability is relatively lower. This interpretation may not be applicable to Bayesian strategies, however.

SUMMARY

The results presented in the previous sections are based on data obtained from prediction analysis, Lord's flexilevel test, Wald's sequential test, Bayesian, and maximum likelihood strategies. MicroCat (Assessment Systems Corporation, 1987) was used for simulations of adaptive testing involving Bayesian and maximum likelihood stategies. Findings of this study may be summarized as follows:

(a) The final predictive probabilities obtained from predictive analysis are significantly correlated with the total scores. These correlations are highly comparable with the correlations between the total test scores and the other strategies.

(b) In terms of the proportion of misclassification into the mastery or non-mastery categories, the predicted adaptive testing perform better than that of sequential probability tests and almost equally well in comparing with Bayesian strategy.

(c) The number of items required is almost the same as the number required by flexilevel test, maximum likelihood and Bayesian strategies. Probably because the number of items in the total test is too small, the benefit of adaptive test could not be demostrated.

(d) There is no effect on the number of misclassification of students into categories when different priors were used in predictive testing. But, in Bayesian decisions, the use of different prior distributions may produce different numbers of misclassifications.

REFERENCES

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press, Cambridge.

Assessment System Corporation (1987). User's Manuel for the MicroCat Testing System. (2nd ed) St. Paul, Minnesota: Author.

Hsu, T.C. and Shermis, M. (1987). Assessing the Psychometric Qualities of a Microcomputerized Adaptive College Mathematics Placement Test. Paper presented at the annual meeting of AERA, Washington, D.C.

Lord, M.F. (1971). The Self Scoring Flexilevel Test. Journal of Educational Measurement. Volume 8, no 3.

Owen, R. J. (1975). A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing. Journal of the American Statistical Association. Volume 70, no 350.

Samejima, F.C. (1977). The Use of the Information Function in Tailored Testing. Applied Psychological Measurement 1, 233- 247.

Wald, A. (1947). Sequential Analysis. John Wiley& Sons, Inc., NewYork.